LEGAL RESEARCH

How to build a more open justice system

Court records are unstructured and costly to access—here's how to fix it

By Adam R. Pah[1,2], David L. Schwartz[3], Sarath Sanga[3], Zachary D. Clopton[3], Peter DiCola[3], Rachel Davis Mersey[4,5], Charlotte S. Alexander[6], Kristian J. Hammond[7], Luís A. Nunes Amaral[2,7]

Modern governments gather information across an extraordinary range of activities and use this information to direct policy. Whether a central bank monitoring inflation or a health agency monitoring disease, these entities typically publicly disclose the information gathered so that their actions can be reviewed and evaluated by others. But in many respects, the justice system is a glaring exception. In the United States, a range of technical and financial obstacles blocks large-scale access to public court records—all but foreclosing their use to direct policy. Yet a growing body of empirical legal research demonstrates that systematic analyses of court records could improve legal practice and the administration of justice. And although much of the legal community resists quantitative approaches to law, we believe that even the skeptics will be receptive to quantitative feedback—so long as it is straightforward, apolitical, and incontrovertible. We offer an example of this kind of feedback as well as a collaborative research agenda to dismantle access barriers to court records and enable the public to analyze them.

Although court records in the United States sit in the public domain, federal courts charge $0.10 per printed page to view any record online (1). Accessing a single case might cost $10 or more. Accessing all cases from a given year would cost millions of dollars (2). To be sure, the federal judiciary releases in-house studies that use federal court records, as well as a database of basic information

about each case, such as the subject matter (e.g., tort, contract, civil rights) and disposition (e.g., settled, transferred, jury verdict) (3). The federal judiciary has steadfastly refused, however, to make the underlying public court records freely accessible.

Selective access is not the approach taken by the rest of the U.S. federal government: Congressional records are freely available at congress.gov. Executive agencies' records are freely available at regulations.gov. It's hard to conceive of a compelling argument for selective access to judicial records that does not apply equally to selective access to congressional records or federal agencies. More to the point, it's hard to conceive of a reason why public records should not generally be accessible to the public.

There are some alternative sources for court records, but barriers to systematic analysis remain. Commercial legal services have directly purchased many court records, but they impose their own fees, prohibit bulk downloads, and thus foreclose systematic analysis even for subscribers. Individual judges and commercial services occasionally grant ad hoc fee reductions for research purposes, but these grants are rare, cumbersome to acquire, limited to subsets of the data, and always come with the condition that the underlying records are not disclosed to the public (4). An open alternative, *Free Law Project*, maintains a crowdsourced repository of free court records, but coverage remains too low to support systematic research.

DATA AND OPENNESS
The lack of access to court records seemingly undercuts any claim that the courts are truly "open" (5, 6). It surely conflicts with researchers' conception of openness. Scientific practice is grounded on a commitment to sharing data and enabling others to replicate findings. But the law's conception of openness is different, a commitment to carrying out public acts in a public space. A scientist might restrict access to a lab and still claim that the research she conducts there is "open." Closed proceedings in a legal setting, on the other hand, are only tolerated in extraordinary circumstances.

Also in contrast to scientific practice, much of the legal profession resists quantitative or evidence-based approaches to improving legal practice and instead prefers to rely on personal experience and professional judgment (7). In a recent Supreme Court case challenging the constitutionality of partisan gerrymandering, Chief Justice John Roberts summarily dismissed empirical approaches

to gerrymandering as "sociological gobbledygook" that any "intelligent man on the street" would denigrate as "a bunch of baloney" (8). Such skepticism is by no means confined to the United States. France, for example, has recently prohibited the publication of any statistical analysis of a judge's or clerk's decisions "with the object or effect of evaluating, analyzing, comparing or predicting their actual or supposed professional practices." Violators face up to 5 years in prison (9).

We believe that these differences help explain why the lack of large-scale access to data is not viewed as a priority—or even as a concern—by much of the legal community. The differences in priorities reflect not just commitments to different values but different conceptions of the same values. Yet, if court records are to be truly accessible and evaluable by the public, the legal and scientific communities must cooperate, and appreciate the values that the other holds dear.

EVALUATING ACCESS TO JUSTICE
Access to justice is a fundamental right and the foundation of any fair and legitimate justice system. But how can one quantify and empirically evaluate this concept? Consider court fees. For a litigant without means, court fees are a substantial barrier to the civil justice system. Anyone who files a lawsuit in federal court must pay a $400 filing fee, along with other costs related to litigation such as formal service of the complaint. Litigants in need can file an application to waive court fees, but there is no uniform standard to review these requests (10). Application forms differ by district. Most ask the applicant to list sources of income, assets, and cash on hand—and then leave the decision to the judge's discretion. Individual judges thus have considerable power over whether to grant or deny access to the justice system.

How do judges exercise this power? This is but one of the myriad questions that is difficult, and arguably impossible, to answer without easy access to structured court records. Even with free access to the data, the answer would be difficult to infer without being able to computationally analyze the text of the court records. In this case, the analysis is straightforward. When a party submits a fee waiver request, the case docket report adds a separate entry for that request, and the textual summary accompanying the entry typically includes some reference to whether the request was granted or denied. We analyzed these entries to compute the grant rate of each federal judge in 2016.

Average grant rates naturally differ among federal districts because cases are not randomly assigned to districts. However, once a case is filed in, say, San Francisco, it is then randomly assigned to one of the judges sitting in the federal district that includes San Francisco. Thus, if all judges reviewed fee waiver applications under the same standard, then grant rates should not systematically differ within districts.

We find, however, that they do (see the figure). At the 95% confidence level, nearly 40% of judges—instead of the expected 5%—approve fee waivers at a rate that statistically significantly differs from the average rate for all other judges in their same district. In one federal district, the waiver approval rate varies from less than 20% to more than 80%.

These findings were recently presented to a group of federal judges who are responsible for amending the rules in their local district. On learning of the inconsistent treatment of fee waiver requests, these judges expressed interest in using our data to improve the decision-making process (11). We count this as an early and encouraging validation of our claim that judges will be especially receptive to quantitative feedback that is straightforward, apolitical, and incontrovertible.

DISMANTLING BARRIERS
Going forward, we believe that the best way to provide the judiciary with quantitative feedback is to develop a forum where individuals can collaborate and build on each other's efforts. With this vision in mind, we propose a three-pronged collaborative research agenda to empower the public to access and analyze court records.

Make court records free
In theory, Congress could make federal records free by repealing the laws that authorize the judiciary to charge for access (12), or the Judicial Conference of the United States (the policy-making body of the federal judiciary) could stop charging fees. Both Congress and the courts have rejected calls to do so. A principal reason, it seems, is money. About 2% of the federal judiciary's budget comes from online record access fees ($145 million in fiscal year 2019). The judiciary is naturally unwilling to forgo this revenue without a commensurate increase from Congress, and Congress, for its part, is unwilling to increase funding. The stalemate persists because not enough judges, members of Congress, and people realize that this is an issue of legitimacy, not just an issue of money.

[1]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [3]Pritzker School of Law, Northwestern University, Evanston, IL, USA. [4]Medill School of Journalism, Northwestern University, Evanston, IL, USA. [5]Moody College of Communication, University of Texas, Austin, TX, USA. [6]Robinson College of Business, Georgia State University, Atlanta, GA, USA. [7]McCormick School of Engineering, Northwestern University, Evanston, IL, USA. Email: amaral@northwestern.edu

To break this impasse, we believe that organizations outside government should directly purchase and publicize court records. The most impactful first step is to make docket reports accessible. A docket report is essentially a lawsuit's table of contents. It lists the case title, presiding judge, subject matter of the suit, and information on the plaintiffs, defendants, and their attorneys. A docket report also gives the date that a document was filed, along with a summary of the document that can be analyzed to extract important features of a case. The data for the figure, for example, were constructed by parsing docket reports,

Inconsistency in judicial fee waiver decisions

Litigants filed 34,001 applications to waive court fees in U.S. federal courts in 2016. For visual simplification, we show only the 294 judges (out of 1742 total) who ruled on at least 35 applications. We would expect 5% of judges to differ from their within-district peers at 95% confidence. Instead, we find that nearly 40% of judges differ.

● 95% confidence interval, not statistically different from zero
● 95% confidence interval, statistically different from zero

Likelihood that judge waives court fees minus likelihood that other judges in the same district waive fees

not the underlying court records. Though docket reports represent only a fraction of all court records, acquiring them will be expensive. The docket reports used in the figure, which cover all cases filed in 2016, cost more than $100,000.

Link data in a knowledge network

Because court records are mostly unstructured text, researchers will need to dedicate extensive time and resources to organizing the data. Documents must be analyzed using natural language processing; entities must be disambiguated; and events, such as

the filing of a fee waiver, must be classified using machine learning. The docket reports should also be linked to external metadata such as information on judges, litigants, and lawyers. By linking court records to outside data sources, individual users can conduct more powerful searches, such as for litigation against big tech firms or for suits currently pending against the federal government.

Although we already have solutions to many of the problems associated with organizing and classifying the data, for many more we will need additional research. For example, it is straightforward to link the presiding judge of each case to outside data on the judge's characteristics such as age, gender, and appointing president. By contrast, to assemble information about litigants and lawyers, researchers will need to make considerable progress on named-entity recognition techniques while protecting litigants' and third parties' privacy. We believe that an open and collaborative platform is the best way to make substantial and rapid progress on these challenges.
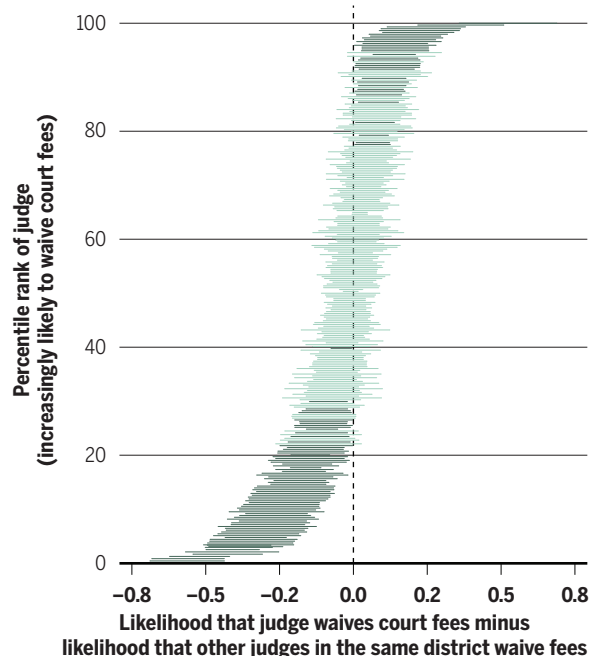
Empower the public

The ultimate goal must be to enable the public to directly evaluate and engage with the work of the courts. To this end, we should create applications that not only support scholars and researchers who may want to analyze the data but also enable members of the judiciary, entrepreneurs, journalists, potential litigants, and concerned citizens to learn more about the functioning of the courts. To support inquiries made by the public, we should develop applications that can process natural language queries such as "What are the most recent data privacy cases?" or "How often do police officers invoke qualified immunity?"

Funding the efforts we propose will be challenging because the cause does not slot nicely into standard philanthropic categories. To carry out our proposals, the academic community should partner with other stakeholders such as nongovernmental organizations, law firms, legal clinics, and other advocacy groups. Indeed, we believe that one of the main reasons why past calls for change failed is because they were not coordinated.

Opening up court records could lead to some flawed or misleading analyses, yet such problems apply to any setting with open data. No one can control what people do with congressional records, federal agency records, census data, etc. Nevertheless, these data are—and should remain—available to everyone. As in any discipline, standards and best practices eventually emerge, and there is already a thriving literature of empirical legal studies. Many scholars have engaged with these data, albeit on a smaller scale. Thus, for the most part, standards and best practices already exist (13).

We believe that the judiciary should be shielded from outside pressures so that it can decide cases according to the law, not the latest poll. But the judiciary also acts on behalf of the public. Its independence must therefore be balanced with commensurate transparency. Ultimately, the judiciary's principal asset is not its annual appropriation from Congress or the revenue generated by access fees, but the public trust. And the most effective way to cultivate this trust—to promote transparency, dismantle barriers to access (14, 15), and build an open knowledge network—is to do it together. ∎

REFERENCES AND NOTES

1. Public Access to Court Electronic Records (PACER), "PACER user manual for CM/ECF courts" (United States Courts, 2019).
2. United States Courts, Federal judicial caseload statistics 2018 (2018); www.uscourts.gov/statistics-reports/federal-judicial-caseload-statistics-2018.
3. W. Hubbard, J. Empir. Leg. Stud. 14, 474 (2017).
4. J. B. Gelbach, Yale Law J. 121, 2270 (2011).
5. A. Bronstad, "PACER fees harm judiciary's credibility, Posner says in class action brief," 25 January 2019; www.law.com/2019/01/25/pacer-fees-harm-judiciarys-credibility-posner-says-in-class-action-brief/.
6. L. Doggett, M. J. Mucchetti, Tex. Law Rev. 69, 643 (1990).
7. H. F. Lynch et al., Science 367, 1078 (2020).
8. Gill v. Whitford, Transcript of oral argument at 38 and 40, no. 16-1161, 138 S. Ct. 1916 (2018).
9. J. Tashea, "France bans publishing of judicial analytics and prompts criminal penalty," ABA Journal, 7 June 2019; www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics.
10. A. Hammond, Yale Law J. 128, 1478 (2018).
11. Owing to the preliminary nature of discussions, the identities of courts and judges are not reported, but Science has confirmed this claim.
12. 28 U.S. Codes §§ 1913, 1914, 1926, 1930, 1932.
13. W. Baude et al., Univ. Chic. Law Rev. 84, 37 (2017).
14. A. Madison, "Team tapped to review PACER amid fee dispute (corrected)," Bloomberg Law, 9 January 2020; https://news.bloomberglaw.com/us-law-week/team-tapped-to-review-pacer-amid-fee-dispute.
15. A. Kragie, "Court transparency bill calls for live audio, free PACER," 2 March 2020; www.law360.com/articles/1249148.